Noname manuscript No.
(will be inserted by the editor)

A Framework to Contest and Justify Algorithmic Decisions

Received: date / Accepted: date

Abstract In this paper, we argue that the possibility of contesting the results of Algorithmic Decision Systems (ADS) is a key requirement for ADS used to make decisions with a high impact on individuals. We discuss the limitations of explanations and motivate the need for better facilities to contest or justify the results of an ADS. While the goal of an explanation is to make it possible for a human being to understand, the goal of a justification is to convince that the decision is good or appropriate. To claim that a result is good, it is necessary (1) to refer to an independent definition of what a good result is (the norm) and (2) to provide evidence that the norm applies to the case. Based on these definitions, we present a challenge and justification framework including three types of norms, a proof-of-concept implementation of this framework and its application to a credit decision system.

Keywords Challenge · justification · machine learning · training dataset · evidence · norm

1 Introduction

The possibility of contesting the results of Algorithmic Decision Systems (ADS) is a key requirement for ADS used to make decisions with a high impact on individuals. This is the case, for example, for decisions made by health professionals, by judges or by bankers. This need is acknowledged, to some extent, by the GDPR, which states that a person who is "subject to a decision based solely on automated processing" has "the right to obtain human intervention

Clément Henin (Corresponding author : E-mail: clement.henin@inria.fr)
Univ Lyon, Inria, INSA Lyon, CITI, Villeurbanne, France
École des Ponts ParisTech, Champs-sur-Marne, France

Daniel Le Métayer
Univ Lyon, Inria, INSA Lyon, CITI, Villeurbanne, France
E-mail: daniel.le-metayer@inria.fr

on the part of the controller, to express his or her point of view and to contest the decision" (Article 22(3)). However, there can be many barriers preventing this right from being exercised, the first one being the practical difficulty to understand the grounds for a decision based on the results of an ADS. To ensure that this right can be effective, we argue that contestability should be supported by appropriate tools. As stated by Mulligan et al., "contestability can support critical, generative, and responsible engagement between users and algorithms, users and system designers, and ideally between users and those subject to decisions (when they are not the users), as well as the public" [28]. Indeed, providing ways to contest a decision can be beneficial in many respects:

- It makes the ADS more effective because it enhances the ability of the human decision maker to detect inappropriate results (suggestions of wrong decisions).
- It makes the ADS more accountable because decisions can be accompanied by justifications.
- It makes the ADS more acceptable from the ethical point of view because it preserves the autonomy of the human decision maker. Indeed, even if the human decision maker does not have any formal obligation to follow the suggestion of the ADS, his/her autonomy is questionable if he/she does not have any possibility to contest it. Because "contestability fosters engagement rather than passivity, questioning rather than acquiescence" [28], it is a key condition to empower human decision makers.

Some authors have already advocated *contestability by design* and analyzed the challenges to be addressed to implement it [2,13]. Opacity is often put forward as a first obstacle to the contestation of ADS based decisions. Indeed, it is difficult to contest the results of a system when information about its logic, operation or input data is too scarce or not intelligible. The possibility to produce intelligible explanations about the results or the overall logic of an ADS is therefore very useful to enhance contestability. However, as stated by Mulligan et. al., providing explanations is not sufficient: "regulatory approaches should seek to put professionals and decision support systems in conversation, not position professionals as passive recipients of system wisdom who must rely on out-of-system mechanisms to challenge them. For these reasons, calls for explainability fall short and should be replaced by regulatory approaches that drive contestable design" [28]. In this paper, we take the same stance and argue that the answer to contestations should be *justifications* and, even though the two words are sometimes used interchangeably, there are essential differences between explanations and justifications.

The word "justification" itself is used with different meanings in the AI literature. In this paper, we propose the following distinctions, which are consistent with T. Miller's characterization of justifications [23]:

- The goal of an explanation is to make it possible for a human being (designer, user, affected person, etc.) to *understand* (a result or the whole system). In contrast, the goal of a justification is to convince that the decision

is *good*. For example, an explanation for a bank loan application rejection could be that the number of outstanding loans is too high. This information helps to understand the logic of the system through a salient feature used by the ADS. The appropriateness of the decision is not questioned. By contrast, a justification of the same decision could be that applications with many outstanding loans have a high probability to lead to credit defaults, which is a risk that the bank wants to reduce. Even if they often support each other, explanations and justifications have different goals: a user can understand the logic leading to a particular result without agreeing on the fact that this result is good; vice versa, he/she may want to contest a result (being convinced that it is bad) without knowing or understanding the logic behind the algorithm.
- Explanations are *descriptive* and *intrinsic* in the sense that they only depend on the system itself. In contrast, justifications are *normative* and *extrinsinc* in the sense that they depend on a reference (or a *norm*) according to which the validity of the results can be assessed. Indeed, in order to claim that a result is good, it is necessary (1) to refer to an independent definition of what a good result is (the norm) and (2) to provide *evidence* that this norm applies to the case. In the above example, the norm is the objective of the bank to reduce credit defaults.

It is important to stress that the notion of norm is central to the challenge-justification dialectic. In general, different types of norms can be applicable to an ADS. These norms can have different sources of legitimacy (legal, ethical, social, economic, etc.) and can be expressed in different ways (e.g. through law or jurisprudence for legal norms). When several norms apply, they may be in tension, or even in contradiction. In some cases, it is possible to rely on priority rules to establish precedence of a norm over another one (e.g., international law usually prevails over domestic law, constitution prevails over ordinary laws, which prevail over decrees, etc.); in other cases, such rules may not exist and the conflicts between them must be solved by a human decision maker on a case by case basis.

Challenges and justifications are dual notions: a challenge can be seen as a statement that a decision is *not good*, supported by evidence, while a justification is a statement that a decision is *good*, supported by evidence. In both cases, evidence refers to a given norm.

As this discussion shows, there can be many different ways to challenge and to justify decisions. Our main contributions in this paper are the following:

- A general definition of the notions of challenge, justification, norm, evidence, statement and argument.
- A generic framework based on the above definitions including a challenge and justification protocol. Three types of norms are considered in this paper; they are inspired by the three main moral theories (virtue ethics, consequentialism and deontological ethics).
- A proof of concept (PoC) implementation of the framework called *Algocate* and its application to a credit decision system.

Section 2 introduces the notions used in the paper illustrated with a case study. Section 3 and Section 4 present respectively the framework and its application to the design of *Algocate*, our challenge and justification system. Section 5 shows examples of our *Algocate* implementation applied to a credit decision system. Section 6 is a discussion of related work and Section 7 concludes with a discussion and suggestions for further research.

2 Challenges and justifications: informal introduction

To illustrate our framework on a concrete example, we introduce an hypothetical bank credit decision system. We assume that this ADS relies on black box machine learning technology. Application files are composed of information about the applicant (age, gender, marital status, revenue, level of education and outstanding loans) and information about the credit itself (amount, duration, interest rate and insurance).

We assume that the norms shown in Figure 1 apply to this case study. These norms have different sources of legitimacy (fundamental rights, sectorial rules, business rules, etc.). However, we do not assume that the legitimacy of a norm is accepted by all parties. For example, in the case study considered here, the bank could refer to internal business rules which are not known or accepted by customers. In such situations, if a party refers to a norm that is not accepted by the other party, the challenge-justification protocol requires the intervention of the human decision maker (called the "arbitrator" hereinafter, even if it does not have to be a arbitrator in the legal sense of the term). The benefit of the protocol in such cases is that a party relying on a norm is compelled to elicit it and to submit it to the approval of the other party (or the arbitrator as a last resort).

1. Compliance with anti-discrimination regulation: gender, ethnicity and country of origin should not have any impact on credit decisions.
2. Banking prudency rule: credit amounts above 20 000 euros must be insured,
3. Bank Objective: the number of credit defaults must be minimized.
4. Bank Policy: ADS decisions must be consistent with the decisions made in the past by human banking agents, subject to compliance with Bank Objective.

Fig. 1 Examples of norms

Figure 2 presents an example of interaction of a user with a justification system in our framework. The statements of the user of the system, which are expressed in natural language, are labelled with *Ui* and the system answers are labelled with *Ai*:

– In *U1*, the user challenges the rejection of his/her credit application and expresses the reasons why he/she believes that his/her application should be accepted. The user is not required to produce any evidence for his/her

U1 My revenue is greater than 50 000 euros and my credit amount is lower than 5 000 euros, so my application should be accepted.
A1 Among the decisions used as reference by the credit decision system, there were 2 000 applications with revenue greater than 50 000 euros and credit amount lower than 5 000 euros and 86 % were accepted (overall average: 70 %).
A2 However, among the decisions used as reference by the credit decision system, there were 260 applications with revenue greater than 50 000 euros, credit amount lower than 5 000 euros and 3 or more outstanding loans (as yours) and 5 % were accepted (overall average: 70 %).
U2 Credit application *CA2* has been accepted. This application has the same attributes as mine, except for the gender. Therefore, my application should be accepted as well.
A3 The initial decision would indeed breach anti-discrimination regulation. Your application must therefore be accepted.

Fig. 2 Example of interaction with *Algocate*

challenge because he/she may not be in a position to do so (typically he/she may not have access to the relevant data).
- In *A1*, the justification system provides evidence supporting the user's challenge. In this case, the evidence is generated from the learning data set of the ADS.
- In *A2*, the justification system provides a justification for the rejection decision and the evidence supporting it. The justification is a refinement of the challenge including an additional attribute, the number of outstanding loans. The evidence is again generated from the learning data set of the ADS. It shows that the number of outstanding loans of the applicant has generally been considered as a strong argument to reject similar applications in the past. In this case, the justification is considered stronger than the challenge because of the very low ratio of accepted applications among reference decisions (5 % of the decisions accepted, which is 65 % less than the average, when the difference for the challenge in A1 is only 16 %), while the number of 260 cases is reasonable. The justification system relies on a strength relationship between arguments to decide if a justification should prevail or not over a challenge. In general, it may be the case that two arguments (justification and challenge) are not comparable. In any case, if two parties disagree, the final decision rests with the arbitrator, the goal of the system being to provide the most valuable information possible to help the latter in this task.
- The interaction could then proceed in different ways depending on the fact that the user accepts or not the norm used to support the decision in *A2*. If he/she does not, then the arbitrator has to decide whether it is acceptable or not. Otherwise, the user can either accept the decision or challenge it in a different way. This is the option followed in Figure 2: *U2* challenges the decision based on anti-discrimination regulation relying on a similar application that has been accepted.

– *A3* is the answer produced by the justification system after verification of the validity of the *U2* challenge. This is the last step of the protocol which, in this case, does not require the intervention of the arbitrator since the parties have agreed on a revision of the decision.

The interactions presented in Figure 2 show that the framework is useful both for users contesting the decisions of the ADS (e.g. individuals affected by the decisions, regulators, or human decision makers who are not sure about the suggestion of the ADS) and for users who want to justify them (usually the operator of the ADS). The justification system is neutral, in the sense that it is designed to find the best arguments (i.e. statements with evidence supporting them) for both parties. This neutrality is of prime importance, given the usual imbalance of powers between individuals who are affected by the decisions and the designers or operators of the ADS.

3 A framework to challenge and justify decisions

Figure 2 illustrates only some of the interactions provided by *Algocate*. In this section, we define more precisely the notions used in this paper and introduce our framework before presenting the *Algocate* system, which is a particular implementation of this framework, in the next section. The framework relies on three notions that were introduced informally in the previous section: statements, norms and evidence. Table 1 provides some examples for each of these notions.

– A statement defines what a decision should be (e.g. "accepted" or "rejected"), according to a party, and the particular aspects of the case (input data of the ADS, e.g. application file) that make him/her believe so.
– A norm is a reference that can be used by a party to support a statement.
– Evidence is used to show that the norm applies to the case.

A triple *¡statement, norm, evidence¿* is called an *argument*, which can be either a *justification* (if the statement supports the decision) or a *challenge* (if the statement contradicts the decision). When conflicting arguments are issued by different parties, it is useful to be able to compare them. To this aim, we will introduce a *strength relation* on arguments.

In the following, we define successively statements (Section 3.1), norms (Section 3.2), and evidence (Section 3.3) before introducing the strength relation (Section 3.4).

We use the following mathematical notations in the sequel. Cases are characterized by tuples $< a_1, \ldots, a_m >$ of m attributes. The set of possible values for attribute a_j is denoted $\mathbf{A_j}$ and $\mathbf{A} = \mathbf{A_1} \times ... \times \mathbf{A_m}$ is the set of all possible cases. The decision function is called $f : \mathbf{A} \rightarrow \mathbb{B}$ with $\mathbb{B} = \{0, 1\}$ (we assume that decisions are binary). We use the notation $x[j]$ to refer to the j^{th} attribute of $x \in \mathbf{A}$. Therefore, $x[j] \in \mathbf{A_j}$. The notation is extended to sets S of cases with $S[j] = \{x[j] | x \in S\}$. For the sake of readability, we use constant names such as "amount" or "duration" rather than indexes in examples. For instance, if the

STATEMENT		
Absolute	$\forall x \in \mathbf{A}, x$[education] $= PhD \wedge x$[outstanding-loans] ≤ 2 $\implies f(x) = 1$	All cases having a PhD and less than two outstanding loans should be accepted
Relative	$\forall x \in \mathbf{A}, (\exists x' \in \mathbf{A}, f(x') = 1 \wedge (x'$[revenue] $\leq x$[revenue] $\wedge x'$[amount] $\geq x$[amount])) $\implies f(x) = 1$	All cases for which there exists an accepted case with lower revenue and greater amount should be accepted
NORM		
Rule (absolute)	$[Rule, \forall x \in \mathbf{A}, x$[outstanding-loans] $\geq 4 \implies f(x) = 0]$	Credit with four or more outstanding loans should be rejected
Rule (relative)	$[Rule, \forall x \in \mathbf{A}, (\exists x' \in \mathbf{A}, f(x') = 1 \wedge (x'$[revenue] $= x$[revenue] $\wedge x'$[education] $= x$[education] $\wedge x'$[age] $\geq x$[age])) $\implies f(x) = 1]$	All cases for which there exists an accepted case with the same education level and revenue and higher age, should be accepted
Objective	$[Obj, \Delta_O, nd]$ (with $x[nd] = 1$ for non-default cases and $x[nd] = 0$ otherwise)	The number of accepted non-default cases should be maximised (the number of defaults should be minimised)
Reference	$[Ref, \Delta_R, d]$ (with $x[d]$ the reference decision for a case x)	The decisions of the ADS should reflect the decisions made in the past by the bank experts
EVIDENCE		
Objective	$[\Delta_O, 1200, 0.06]^{(i)}$ where Δ_O is the historical database with the objective attribute $x[o]$	The non-default rate for the 1200 historical credits corresponding to the statement is 93 % (overall average: 87 %)
Reference	$[\Delta_R, 500, 0.28]^{(ii)}$ where Δ_R is the reference database with the reference attribute $x[d]$	Among the 500 past experts decisions corresponding to the statement, 95 % were accepted (overall average: 67 %)

Table 1 Examples of statements, norms and evidence. (i) The value 0.06 is the result of the difference 0.93 - 0.87. (ii) The value 0.28 is the result of the difference 0.95 - 0.67.

third field of a case x represents the amount of a credit, we write $x[amount]$ to denote it (with $amount = 3$). The case under consideration is called $x_s \in \mathbf{A}$.

3.1 Statements

As suggested in the previous section, statements involve two pieces of information:

- The decision that should be made, according to the issuer of the statement (1 for "accepted" or 0 for "rejected") and
- The property of the case that argues in favour of this decision, according to the issuer of the statement.

In mathematical terms, statements are therefore defined as follows:

$$\forall x \in \mathbf{A}, C(x) \implies f(x) = \delta \tag{1}$$

with $\delta \in \mathbb{B}$ the decision supported by the issuer and $C(x)$ the property supposed to justify this decision. A statement is relevant for a case x_s only if $C(x_s)$ is true, which will be assumed thereafter. If $\delta \neq f(x_s)$, the goal of the statement is to contest the decision. We call it a *challenging statement*. Vice versa, if $\delta = f(x_s)$, the goal of the statement is to support the decision. We call it a *justifying statement*.

In general, $C(x)$ could involve comparisons of x with any number of other cases. For the sake of simplicity, we consider only two options here: conditions involving zero or one other case. The definitions can be easily generalized to any number of other cases[1].

The first type of statements, called *absolute statements*, does not refer to any other case. Examples of absolute statements appear in the first interaction step ($U1$) of Figure 2 and in the first line of Table 1. The general form of condition $C(x)$ for absolute statements is the following:

$$C(x) = (x[i_1] \diamondsuit_1 v_1) \wedge \ldots \wedge (x[i_k] \diamondsuit_k v_k) \tag{2}$$

[1] We have not encounetred any practical situation in which such generalization would be useful, though.

with $i_p \in \{1, \ldots, m\}$, $\diamondsuit_p \in \{=, \leq, <, \geq, >\}$ for p in $\{1, \ldots, k\}$ and $v_p \in \mathbf{A_{i_p}}$. The first line of Table 1 provides an example of condition expressed in this syntax.

In contrast, *relative statements* involve a comparison with another case x'. The second user interaction $(U2)$ of Figure 2 and the second line of Table 1 are examples of relative statements. More formally, condition $C(x)$ for relative statements has the following form:

$$\begin{aligned} C(x) = \exists x' \in \mathbf{A}, f(x') = \delta \quad \wedge \\ (x'[i_1] \diamondsuit_1 x[i_1]) \wedge \ldots \wedge (x'[i_k] \diamondsuit_k x[i_k]) \end{aligned} \tag{3}$$

with for all p in $\{1, \ldots, k\}$, $i_p \in \{1, \ldots, m\}$, $\diamondsuit_p \in \{=, \leq, <, \ll, \geq, >, \gg\}$. Operators $\ll$ and $\gg$, which stand for, respectively, "much less than" and "much greater than", are defined as follows: $x \ll y \Leftrightarrow x + k \leq y$ and $x \gg y \Leftrightarrow x \geq y + k$ with k an application-dependent parameter. The statement corresponds to a situation in which another case x' is associated with a decision δ and the relation between the two cases would justify that the same decision is made for x. The second line of Table 1 provides an example of condition expressed in this syntax.

3.2 Norms

As discussed in the introduction, in order to challenge or to justify a decision it is necessary that a statement is backed by an applicable norm. We consider three types of norms here, called respectively *rule-based*, *objective-based* and *reference-based* norms, which correspond to three typical ways to support a challenge (or a justification). Interestingly, these three modes also reflect the approaches followed by the three main families of moral theories (respectively deontological ethics, consequentialism and virtue ethics). Other types of norms can be easily added to the framework, provided that their meaning is defined precisely, as done in this section, and their strength is characterized as done in Section 3.4.

Rule-based norms

Examples of rule-based norms expressed in an informal way can be found in the first two lines of Figure 1. As the name suggests, a rule-based norm is defined by a fixed rule. Formally speaking, these rules can be expressed as:

$$[Rule, Def]$$

with $Rule$ a flag defining the type of the norm and Def its content. Def is expressed in the same language as statements, that is, definition (1) of Section 3.1 with C defined by equation (2) if the norm is expressed in terms of a single case (absolute rule) or (3) if it involves a second case (relative rule). Lines 3 and 4 of Table 1 show respectively an example of absolute rule (business rule) and an example of relative rule.

This type of norm is common, *inter alia*, in law (regulations, directives, acts, contractual rules, etc.) and business (sectorial rules, corporate rules, procedural rules, etc.). In terms of moral theories, it is also the spirit of deontological ethics[2] which relies on moral obligations (such as "Thou Shalt Not Tell Lies") that must be followed by any rational agent [32].

Objective-based norms

The second way to define a norm consists in using measurable objectives that can be used to assess decisions. The third line of Figure 1 shows an informal example of objective-based norm. In formal terms, an objective is expressed as an attribute that should be maximised[3]. In Figure 1, this attribute is the number of non-default cases for Bank Objective.

Formally speaking, objective-based norms can be expressed as

$$[Obj, \Delta_O, ob]$$

where Obj is a flag defining the type of the norm, Δ_O a database containing the values $\Delta_O[ob]$ of the objective attribute ob. Line 5 of Table 1 shows an example of objective-based norm.

Objective-based norms are common in business and organizations in general. In terms of moral theories, they can be related to consequentialism, which, as stated by Mark Timmons "explains the deontic status of actions and other items of moral evaluations entirely in terms of the values of the consequences of actions and other items being morally evaluated". Even if we do not restrict the scope of norms to moral issues, as shown in the examples, objective-based norms follow the same approach as consequentialism in the sense that, rather than relying on fixed rules (as rule-based norms and deontological ethics), they focus on the assessment of the impact of the decisions.

Reference-based norms

The third type of norm is illustrated by Bank Policy in Figure 1. This type of norm is applicable when reference data Δ_R about past decisions is available. The principle in this case is that the decisions are justified if they are consistent with past decisions. The implicit assumption is that the reference data is valid or legitimate, in the sense that it can be used as a model for future decisions. This assumption can actually be disputed when a decision is challenged, as discussed in Section 4.

Formally speaking, reference-based norms can be expressed as

$$[Ref, \Delta_R, d]$$

where Ref is a flag defining the type of the norm, Δ_R a database containing the values $\Delta_R[d]$ of the decision attribute d. Line 6 of Table 1 shows an example of reference-based norm.

[2]The main representative of this school of thought is Kant's moral theory.

[3]Of course, it is also possible to express minimisation by considering the opposite of the relevant attribute.

This type of norm is common in law (use of previous cases or jurisprudence). In terms of moral theories, they can be related to virtue ethics, which is introduced as follows by Mark Timmons: "We often look to others as models for the type of person we would like to be because we think they possess certain admirable character traits." Indeed, we can see the reference data as a model of "good behaviour" (here a model for "good decisions"), the goal of the ADS being to mirror as closely as possible the behaviour of this model. When an ADS relies on supervised machine learning, the learning data set can obviously be used as the reference data.

3.3 Evidence

In an argument ¡*statement, norm, evidence*¿, the *evidence* component shows how the *statement* is supported by the *norm*. Evidence can take different forms depending on the type of the norm. Actually, the main difference is between:

- On the one hand, rule-based norms, which do not rely on databases: they either support or do not support a statement (binary stance).
- On the other hand, objective-based norms and reference-based norms, which rely on databases and can be supported by quantitative evidence. A statement can be more or less supported by such norms.

The case for rule-based norms is simple since these norms are expressed in the same language as statements. A statement:

$$\forall x \in \mathbf{A}, C_s(x) \implies f(x) = \delta_s \tag{4}$$

is supported by a rule:

$$\forall x \in \mathbf{A}, C_r(x) \implies f(x) = \delta_r \tag{5}$$

if and only if:

$$\delta_s = \delta_r \wedge \forall x \in \mathbf{A}, (C_s(x) \implies C_r(x)) \tag{6}$$

which can be derived using a simple inference system comparing term by term the $(x[i_j]\diamondsuit_j v_j)$ components of C_s and C_r.

Evidence involving objective-based norms and reference-based norms relies on two numerical values called respectively the "coverage" and the "deviation". If the statement is defined by:

$$\forall x \in \mathbf{A}, C(x) \implies f(x) = \delta \tag{7}$$

and Δ is the relevant dataset (either Δ_O or Δ_R), we first define $\Delta_{|C}$, the subset of cases matching the condition of the statement:

$$\Delta_{|C} = \{x \in \Delta | C(x)\} \tag{8}$$

The coverage $\gamma(\Delta, C)$ is defined as follows:

$$\gamma(\Delta, C) = card(\Delta_{|C}) \tag{9}$$

with $card(S)$ the cardinal of a set S.

For objective feature ob, the deviation $\mu(\Delta, C, \delta)$ is defined as follows:

$$\mu(\Delta, C, \delta) = (2\delta - 1)(\overline{\Delta_{|C}[ob]} - \overline{\Delta[ob]}) \tag{10}$$

with $\overline{S}$ the average of the values of the set S. The deviation μ measures the difference of averages between the subset characterized by C and the whole population. The factor $(2\delta-1)$ is justified as follows. If the difference is positive (objective higher in $\Delta_{|C}$) then the evidence supports statements such that $\delta = 1$ and negates statements such that $\delta = 0$. Vice versa, if the difference is negative (objective lower in $\Delta_{|C}$) then the evidence supports statements $\delta = 0$ and negates statements with $\delta = 1$. For the sake of readability, it is preferable to show to the user both the subset average $\overline{\Delta_{|C}[ob]}$ and the population average $\overline{\Delta[ob]}$ as done in Table 1. The deviation for reference-based norms is defined in the same way.

The full definition of evidence is the following:

$$[\Delta, \gamma(\Delta, C), \mu(\Delta, C, \delta)] \tag{11}$$

with Δ the relevant dataset (Δ_O or Δ_R). The last two lines of Table 1 show examples of evidence for an objective-based norm and a reference-based norm. As suggested above, data-based evidence can be more or less supporting. For instance, in the last line of Table 1, the statement characterizes a subset of size 500 with an average number of accepted applications 28 % higher than the whole population (95 % - 67 % = 28 %). In this case, the argument is strong because both the coverage (500) and the deviation (0.28) are high, but it is not the case for the penultimate line of Table 1 in which the deviation is only 0.06. We discuss further the notion of strength in the next section.

We consider only well-formed arguments here, that is, arguments $< S, N, E >$ such that E supports, to some extent, S. More precisely, if N is a rule-based norm, then argument $< S, N, E >$ is well-formed only if Condition (6) is met. If N is a data-based norm, it is well-formed only if $\gamma(\Delta, C) \neq 0$ and $\mu(\Delta, C, \delta) > 0$.

3.4 Strength relation

When two parties disagree about a decision and each party provides an argument to support his/her position, it is important to be able to compare these arguments. To this aim, we define a preorder relation[4] $\geqslant_a$ between arguments. This relation (hereafter "strength relation") is defined in terms of the components of the arguments:

$$[S, N, E] \geqslant_a [S', N', E'] \Leftrightarrow (N \geqslant_n N') \vee ((N = N') \wedge (E \geqslant_e E')) \tag{12}$$

In other terms, an argument A is stronger than an argument A' in two cases:

[4]Binary relation that is reflexive and transitive.

- A relies on a stronger norm than A' or
- A and A' rely on the same norm but A's evidence is stronger than A''s evidence.

This generic definition can be complemented, on a case by case basis, depending on the context[5].

The strength relation $\geqslant_n$ between norms is essentially domain-dependent and it has to be defined for each application. As an illustration, in the example of Figure 1, the bank has specified an explicit strength relation between two norms (Bank Objective $\geqslant_n$ Bank Policy). In addition, the first norm, which is a fundamental right, is stronger than all the others. It should be clear that the strength relation is partial: there are situations in which two applicable norms are not comparable. We come back to this issue in the next section.

As far as $\geqslant_e$ is concerned, it is only relevant for data-based evidence, since rule-based evidence is not quantitative. A simple way to define it for objective-based norms could be the following:

$$[\Delta, \gamma, \mu] \geqslant_e [\Delta, \gamma', \mu'] \Leftrightarrow (\gamma \geq \gamma' \wedge \mu \geq \mu') \tag{13}$$

Large coverages lead to stronger arguments because they improve the statistical significance of the evidence. The strength of the evidence also grows with μ because higher values of μ correspond to more supportive objective values or reference decisions. This definition of $\geqslant_e$ is intuitive but it is conservative in the sense that it leads to a great number of incomparable pieces of evidence. A more sophisticated definition is proposed in the next section.

4 *Algocate*: a challenge and justification system

The framework introduced in the previous section can be instantiated in different ways to build a challenge and justification system. In this section, we sketch the main choices made in the design of *Algocate*, our proof-of-concept implementation. We first describe in Section 4.1 the main steps of the interaction protocol, which shows the functionalities provided by the system. Then, we define more precisely the strength relation used in *Algocate* in Section 4.2 and its use for the generation of statements in Section 4.3 .

4.1 The *Algocate* protocol

An *Algocate* session is always associated with a decision, which is called the *initial decision* in the sequel. The user can be any stakeholder or party concerned by the ADS (designer, operator, human decision maker, person affected by the decision, auditor, etc.) and his/her motivation can be varied, for example to find arguments to support the initial decision, to contest it or to enhance his/her trust. We also assume that, in case of disagreement, final

[5] See Section 4 for an example.

decisions are always taken by a human agent (called the arbitrator in this paper). Last but not least, we assume that a set of norms and evidence have been supplied to *Algocate* by the stakeholders and/or by independent third parties (e.g. regulation bodies). However, no assumption is made neither about the comprehensiveness of this initial set of information nor about the fact that all users will necessarily accept these norms. As shown below, this set of information may evolve, in particular to take into account the verdicts of the arbitrator.

1. The interaction with *Algocate* starts with an initial statement (called the *user statement*) by a user. If the user has a specific norm in mind to support his/her statement, he/she can also provide it, but this information is not mandatory. In the following, we consider the most common (and most complex) situation in which the only input to the system is a user statement.
2. *Algocate* analyses the user statement and searches for appropriate norms and evidence to generate the strongest well-formed argument(s) supporting this statement. Since the strength relation is not total, some arguments may not be comparable. In such cases, *Algocate* returns several arguments. The strength relation used by *Algocate* is defined in Section 4.2.
3. The next step for *Algocate* is to try to reinforce the user statement to find stronger arguments. At this stage, *Algocate* adopts a neutral position and considers both arguments supporting the initial decision and arguments challenging it. *Algocate* returns the strongest well-formed arguments based on these new statements (called *generated statements*). The generation of statements is described more precisely in Section 4.3.

Several options are possible for the user at this stage. The most interesting situation is the case of the user whose goal is to challenge the initial decision:

- If *Algocate* has generated strong evidence supporting the statement of the user (in Step (2) or Step (3)), then he/she can provide this evidence to the arbitrator to request a revision of the initial decision.
- If *Algocate* has generated stronger evidence against the statement of the user (in Step (3)), then the user can either be convinced that the initial decision is legitimate or not. In the first case, the protocol stops since the disagreement has been solved. In the second case, the first option for the user is to try to challenge the decision on a different basis (with a new statement), trigerring a new iteration of the protocol. The second option, if he/she believes that the argument generated by *Algocate* is not legitimate, is to submit it to the arbitrator. This can be the case, for instance, if the argument relies on a norm which is not accepted by the user. Typical examples can be biased reference data or norms corresponding to corporate rules that are not known by customers. If the arbitrator confirms that the norm is not acceptable, the impact of his/her verdict goes beyond the decision challenged by the user: the operator of the ADS must modify the system to correct the situation (and the set of norms used by *Algocate* should be updated accordingly).

4.2 Strength relation

The strength relation used to compare arguments should meet two criteria: it should reflect the intuition of the parties (and the intuition of the arbitrator) and it should make it possible to compare all (or most) arguments. These two criteria can be in tension, as shown in Section 3.4 which introduces a simple and intuitive relation that leaves many arguments incomparable. To solve this tension, *Algocate* relies on a *score function* t measuring the strength of data-based evidence such that:

$$[\Delta, \gamma, \mu] \geqslant_e [\Delta, \gamma', \mu'] \Leftrightarrow t(\Delta, \gamma, \mu) \geq t(\Delta, \gamma', \mu') \tag{14}$$

$$t(\Delta, \gamma, \mu) = \mu\sqrt{\gamma} \tag{15}$$

Intuitively, evidence is strong if the deviation is high, meaning that condition C has a strong impact on the average. However, if the size of $\Delta_{|C}$ is too small, the deviation could be high only by chance. For instance, evidence involving a set of only two reference cases should be considered weaker than evidence relying on a subset of one hundred reference cases (if their deviations are close). To take this factor into account, it is a common practice to compare the average value of $\Delta_{|C}[o]$ with the expected average value of a random drawing of the same size. From the law of large numbers, we know that the expected standard deviation of this randomly drawn subset is proportional to $1/\sqrt{\gamma}$. Definition (15) amounts to the well-known *student's t-test*[6] and it can be converted into a *p-value*[7]. Even though Definition (15) is rather intuitive and has good statistical properties, it is possible to opt for different versions of strengh in *Algocate* (e.g. relying on Shannon entropy). We do not discuss them further in this paper for the sake of conciseness.

4.3 Generation of statements

For a neutral challenge and justification system, it is not sufficient to generate evidence to support the statement issued by the user. As shown in Figure 2, the fact that this statement is supported by some evidence does not mean that a stronger argument cannot be found to support either the same position or the opposite.

The objective of *Algocate* statement generation procedure is illustrated by Figure 3. The left side of the figure (a) shows the historical data set Δ with the initial decision in red and the statement of the user (stating that the decisions in the orange hatched area, which represents $\Delta_{|C}$, should be 1). This statement is weakly supported by the dataset. The middle part of the figure (b) shows a stronger argument supporting the same conclusion (with an additional condition represented by the horizontal bar). However, the right side of the

[6] Up to a constant involving the standard deviation of $\Delta[o]$.

[7] The p-value can be interpreted as the probability that the observed deviation is coincidental

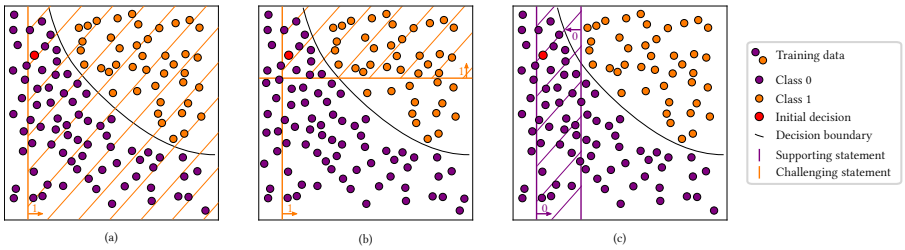

Fig. 3 Visual representation of the generation of absolute statements for reference-based norms. Plots represent the reference (training) data. The two classes are represented in orange and purple. (a) The initial statement of the user (orange vertical line) states that all decisions in the orange hatched area should be 1. It is not strongly supported by the training data (the orange hatched area contains many purple points). (b) Optimal statement supporting the position of the user : it states that all decisions in the orange hatched area should be 1. Strong evidence supports this statement as the ratio orange / purple in the hatched area is high. (c) Optimal statement against the position of the user: it states that all decisions in the purple hatched area should be 0. Strong evidence supports this statement as the ratio purple/orange is 1.

figure (c) shows an even stronger argument against the statement of the user (represented by the two vertical bars). This is indeed the strongest argument according to the strength relation $\geq_e$ defined in the previous section, which is consistent with the intuition since the size of the selected set is approximately similar but the ratio of negative decisions is much higher than the ratio of positive decisions in (b).

More precisely, the goal of the search procedure is to find conditions C^* and conclusion δ (0 or 1) such that:

1. the generated statement includes the initial decision x_s: $C^*(x_s)$ is true,
2. the generated statement strengthens the initial statement: $\forall x, C^*(x) \implies C(x)$,
3. the evidence supporting this statement is maximal: $t(\Delta, \gamma^*, \mu^*)$ is maximized.

with $\gamma* = \gamma(\Delta, C^*)$ and $\mu* = \mu(\Delta, C^*, \delta)$.

Coming back to Figure 3, we can see that all hatched areas include the initial decision (red point) and strengthen the initial statement (the vertical line in (a) is reused in (b) and (c)). Also, we see intuitively that a strong statement should cover as many corroborating data points as possible (many orange points in Figure 3 (b) and purple points in 3 (c)).

The benefit of generating statements that strengthen the argument of the user is to ensure that the answers of the system take into account the concern of the user. This customization makes the interaction more constructive and insightful. Of course, it is always possible, for any party, to start a new interaction on a completely different basis (with a different statement), as illustrated by U2 in Figure 2. Further examples of statement generation are presented in Section 5.

Technically speaking, the search procedure considers all possible norms in decreasing strength order. Each step takes as inputs (1) the initial decision

x_s, (2) the user statement which contains two pieces of information (C and δ) and (3) the database Δ corresponding to the norm. For absolute statements, it outputs a set of triplets $\{(a_i^*, \diamondsuit_i^*, v_i^*), i = 1...K\}$[8]. The output statement is obtained by concatenating these triplets to the initial statement:

$$C^* = C \wedge a_1^* \diamondsuit_1^* v_1^* \wedge ... \wedge a_K^* \diamondsuit_K^* v_K^* \tag{16}$$

With this definition of C^*, the coverage of the generated statement is $\gamma(\Delta, C^*)$ (the number of points in the hatched areas of Figure 3 (b) and (c)). The goal of the search procedure is to find the set of triplets resulting in the greatest value of $t(\Delta, \gamma, \mu)$. The global search objective can be written:

$$\max_{(a_i', \diamondsuit_i', v_i'), i=1...K} t(\Delta, \gamma', \mu') \tag{17}$$

As there is no analytical solution to this maximization problem, to the best of our knowledge, it is implemented by an exhaustive search strategy. Furthermore, as the complexity of the problem is exponential in K, a greedy search algorithm is used to find an approximate solution in a reasonable time. At each step, all possible triplets $(a_i', \diamondsuit_i', v_i')$ such that $x_s[a_i'] \diamondsuit_i' v_i'$ are considered and the triplet leading to the greatest value of $t(\Delta, \gamma', \mu')$ is selected. The iteration stops when the p-value associated with the t-test of the best triplet is above a given threshold[9].

5 *Algocate* at work

In this section, we provide some examples of use of our proof-of-concept implementation of the *Algocate* system to illustrate the benefits of the framework and the feasibility of the approach.

We use as a case study a publicly available dataset: the German credit dataset[10] called Δ_G in the sequel. It contains information about credit applications (credit amount, savings status, etc.) and the conclusions of the bank experts (low risk or high risk). Table 2 shows the values of attributes used in the examples presented in this section. We trained a random forest classifier on Δ_G to build an ADS predicting the conclusions of the bank experts.

We use three norms N_1, N_2 and N_3 for this case study, with $N_1 \geq_n N_2 \geq_n N_3$:

- N_1 (rule-based norm) :
  $\forall x \in A, x[\text{savings\_status}] = \text{“no savings”} \implies f(x) = 0$
- N_2 (reference-based norm): Δ_G should be used as a reference dataset with *decision* the reference attribute.

[8] Set of pairs $\{(a_i^*, \diamondsuit_i^*), i = 1...K\}$ for relative statements. We do not discuss further relative statements here, as they are implemented in a very similar way.

[9] The current implementation of *Algocate* uses the value 0.2, but this value is a parameter that can easily be ajusted.

[10] https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

– N_3 (objective-based norm): Δ_G should be used as a objective dataset with the opposite of the *duration* attribute used as the objective. The strategy of the bank is to favour credits with short durations to reduce the risk of defaults and to foster short term profits.

The value of an objective attribute cannot be known at the time of the decision as it should be a consequence of the decision used to assess its impact. Therefore, for the sake of this example, we assumed that the *duration* attribute was not known at the time of the decision (although it is included in the German credit data set) in order to use it as an objective attribute. In this hypothetical credit decision system, the credit duration is supposed to be flexible and the borrower can adapt the payments to his/her reimbursement capacities.

Operators $\ll$ and $\gg$ ("much less than" and "much greater than", as introduced in Section 3.1) are defined with k equal to twice the estimated standard deviation of the corresponding attribute. The interactions shown in Section 5.1 and Section 5.2 were generated using our PoC implementation of *Algocate*. As in Figure 2, the statements of the user, which are expressed in a restricted natural language[11], are labelled with *Ui* and *Algocate* answers are labelled with *Ai*. The computation time on a laptop was of the order of few seconds for each example.

	Example 1 (5.1) case A	Example 1 (5.1) case A'	Example 2 (5.2) case A
checking_status	<0	<0	0<=X<200
residence_since	4	1	2
number_dependent	1	2	1
credit_amount	2578	7629	2762
age	55	46	25
install_commitment	3	4	1
savings_status	<100	100<=X<500	no savings
decision	0	1	0

Table 2 Attribute values of the cases used in the examples. The checking_status and saving_status attributes refer respectively to the status of the checking and saving accounts. The install_commitment attribute is a categorization (from 1 to 4) of the ratio of the repayment instalments and disposable income (higher values being associated with higher ratios). The values of the decision attribute are 0 (for high risk) and 1 (for low risk).

5.1 Example 1

This first example illustrates the ability of *Algocate* to generate "optimised" statements. Here, the user is a customer whose credit application file A has been rejected and who wants to challenge the decision. The user happens to know another application A' that has been accepted and uses it in his/her statement (which is therefore a relative statement). The values of the features corresponding to user A and user A' are presented in Table 2. For the sake of this first example, we consider only reference-based decisions. The interactions

[11] Transcript of the abstract syntax used in *Algocate*.

shown in Figure 4 follow the three steps of the protocol presented in Section 4.1: statement of the user (U1), argument generated by *Algocate* for this statement (A1) and a stronger argument (based on a new statement generated by *Algocate*) supporting the same position. At the end of the interaction, the customer is in a better position to challenge the decision (for example with the bank or in court).

U1 A' has been accepted. A' has a checking_status value equal to A and a residence_since value much lower than A. Therefore, A should be accepted as well.

A1 Among the decisions used as reference by the credit decision system, there were **413** reference decisions with the same checking_status value, and a much greater residence_since value than an accepted case and **66 %** of them were accepted (population average 70 %).

A2 Among the decisions used as reference by the credit decision system, there were **157** reference decisions with the same checking_status value, a much greater residence_since value, a much lower number_dependent value and a greater age value than an accepted case and **88 %** were accepted (population average 70 %) (strongest generated argument).

Fig. 4 Example of interaction with *Algocate* for a rejected credit application A wishing to challenge the model relatively to another file A'. Evidence shown in A1 is weak (only 66 % of accepted reference decisions). Evidence shown in A2 is stronger (88 % of accepted reference decisions).

5.2 Example 2

The second example shows how *Algocate* can be used by different parties to challenge or to support a given decision. Here, the first party is a customer whose application has been rejected and who disagrees with the decision and the second party is a banking agent who is asked to justify the decision.

The customer issues a challenging statement and obtains an argument with strong objective-based evidence (Figure 5). Let us assume that the customer decides to contest the decision with this argument in front of the bank. To support the decision of the system, the agent issues his/her own statement (Figure 6). *Algocate* finds rule-based evidence to support this statement. Assuming that this argument is stronger than the customer's argument generated in Figure 5, the customer can either accept the decision (for example, if he/she realizes that N_1 is imposed by law) or decide to challenge it (for example, if N_1 is an internal rule of the bank that does not reflect common practice). If the verdict of the arbitrator is in favour of the customer, the operator of the ADS should modify in to take this decision into account and the norm database of Algocate should be updated accordingly.

U1 *A* credit_amount value is lower than 4000 and age is greater than 25, so *A* should be accepted.
A1 Among the decisions used as reference by the credit decision system, there were **602** decisions with credit_amount value lower than 4000, and age value greater than 25. The average credit_duration among them is **17** months (population average 20 months).
A2 Among the decisions used as reference by the credit decision system, there were **116** decisions with credit_amount value lower than 2764, age value greater than 25, and install_commitment value lower than 2. The average duration among them is **12** months (population average 20 months) (strongest generated argument).

Fig. 5 Interaction initiated by a customer challenging the initial decision. In A1 and A2, evidence supports the user statement as the final duration of the credit is below the population average, which corresponds to objective N_3.

U1 *A* savings_status value is equal to "no savings", so *A* should be rejected.
A1 This statement is supported by the rule-based norm N_1.

Fig. 6 Interaction initiated by the banking agent to justify the decision

6 Related Work

In the field of explainable AI, the distinction between explanations and justifications is sometimes blurred. In some papers [19], justifications are seen as ways to make understandable the inner operations of a complex system, which would be called white-box explanations according to our definitions. Similarly, [9,8] make a distinction between "how did the system arrive at the prediction" and "why should we believe the prediction is correct" but they do not refer to any external norm, and therefore provide explanations (based on narrative roles explaining the effect of the features of the input data used by the system) rather than justifications in our sense. Another distinction is made by Klass and Finin [15] based on the intention, which should be to "produce knowledge in the hearer" for explanations and "to affect the beliefs of the hearer" for justifications. From a different perspective, [10], introduces a classification of justifications in machine learning related to the performances (accuracy) of the systems.

A series of works [7,23,27] refer to justifications as ways of ensuring that a decision is good (in contrast to understanding a decision), which is in line with the approach followed in this paper and definitions of explanations and justifications in philosophy [3]. Regarding the extrinsic nature of explanation, [30] distinguishes between justifications and explanations based on the origin of the information they refer to: explanations describe how the system works while justifications use domain knowledge to show that decisions are correct. The normative nature of justifications has also been pointed out in the field of intelligent systems [17]: "an intelligent system exhibits justified agency if it follows society's norms and explains its activities in those terms". In [18], the

authors qualify explanations as "unjustified" when there are not supported by training data, which is related to our notion of justification with reference-based norms (Section 3.2). However, in this context justifiability applies to explanations rather than to the decisions themselves.

The term "justification" is sometimes used in the field of autonomous agents to refer to motivations to perform a specific action. The main difference with our approach is that norms are made available to autonomous agents to make their decisions. For example, the framework introduced in [20] relies on norms (called "principles") inferred from ethical judgments using inductive logic programming. In a nutshell, a value-driven agent refers to an ethical preference ordering of the actions before making any decision. Justifications of the decisions can be built using assumption-based argumentation with deductive rules representing the acceptance of different ethical consequences. It is worth noting that it is appropriate to present justifications as a type of explanation in this context since norms are internalised. Our approach is complementary to this trend of work in the sense that it is not limited to norms expressed in terms of rules and situations in which all choices can be automated.

Finally, closer to our approach, [21] criticizes "anormative" explanations and calls for the explicit definition of algorithm "goals" that should be understandable. These goals being defined, a decision is justified when the "evidence" that it meets the goals can be provided. However, unlike the framework proposed in this paper, goals must be defined in advance and included in the design objectives of the system. Also the possibility to contest a decision based on these goals is not mentioned. In the same spirit, but at another level, [31] uses a refinement structure to provide justifications about the decisions made during the design of the system.

The need to design systems that are contestable has been pointed out by several authors recently. For instance [2] calls for systems that are "contestable by design". In [29], the authors show the importance of being able to challenge algorithmic decisions or recommendation systems in the field of medicine.

The interest for more interactive machine learning systems manifests in a variety of ways [1]. The need to conceive explanations as an interactive process has been argued by several authors [24], [26]. The "human-in-the-loop" approach leverages on human feedback during the training process to obtain more accurate classifiers [16]. A lot of work has also been done on argumentation and dialog games [4,5,22] but the focus in these areas is generally the logical structure of the framework to express and to relate arguments or the protocol to exchange arguments. In contrast, we take an empirical approach to assess challenges and justifications and we consider a very basic protocol in this paper (Section 4.1).

More closely related to our work, [14] relies on "debates" between two competing algorithms exchanging arguments and counterarguments to convince a human user that their classification is correct. However, the goal of this work is to "align an agent's actions with the values and preferences of humans", which is seen as a "training-time problem". Our objective in this paper was different

but an interesting avenue for further research could be the application of our approach to design or to improve an ADS.

To the best of our knowledge, none of these contributions has led to the proposal of a framework or a tool to challenge and to justify decisions comparable to the work presented in this paper.

7 Conclusion

The ultimate goal of the work presented in this paper is to improve the integration of algorithmic decision tools in the overall decision making process and ensure that they can be used for the benefit of people relying on them or affected by their decisions. We emphasize that the framework presented here is not intended to replace a human decision maker but to put him/her in the best position to make a decision. It can also be useful to better justify decisions and to reconcile the viewpoints of the parties.

A system like *Algocate* is also a contribution to accountability, which has been pointed out as a key and complex issue for ADS [6,11]. For example, [6] states that "anyone deploying algorithmic models inevitably also engages a set of normative principles (at least implicitly)" and "accountability involves the provision of reasons, explanations and justifications and this ought to involve drawing out these implicit epistemic and normative standards". As such, *Algocate* could complement proposals such as *model cards* [25] (which can be used to provide information about a machine learning model, such as its intended use, metrics to assess it potential impacts, training data, evaluation data, quantitative analyses, etc.) and *datasheets for datasets* [25] (which can be used to describe the characteristics of a dataset) in order to provide an interactive and easily accessible form of accountability.

We believe that this kind of system is most welcome in situations in which the stakes are high (e.g. in the healthcare or justice sectors) and the requirements of the ADS are not entirely formalized or cannot be guaranteed by design. This may be the case for various reasons, for example when requirements involve ethical aspects, or legal aspects leaving room for interpretation, or evolving constraints (e.g. procedures or regulations), or technical aspects which are not prone to formal definitions (e.g. image or text analysis). As an illustration, the legal constraints applicable to a given system are rarely formalized before its development[12]. A system like *Algocate* makes it possible to provide, independently from the design of the ADS and incrementally, the material (norms and evidence) useful to justify decisions. We hold that this contribution is a first step to address the call expressed in [12] "for tangible action to move from high-level abstractions and conceptual arguments towards applying ethics in practice and creating accountability mechanisms."

As far as further work is concerned, *Algocate* should be tested through randomized user studies involving different types of users in order to prove its

[12]In any case, regulations cannot be entirely formalized.

usability in real life. We plan to carry out these experiments in the near future in collaboration with partners in the public sector (tax authorities) and the private sector (insurance companies).

As stated in the introduction, our framework works in a black-box mode, in the sense that no assumption is made about the internals of the ADS. As a result, it may also be useful to justify or contest decisions made by human beings (without the help of an ADS). Further work is needed to assess the relevance of *Algocate* in this context.

References

1. Abdul, A., Vermeulen, J., Wang, D., Lim, B.Y., Kankanhalli, M.: Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18, p. 1–18. ACM Press (2018). DOI 10.1145/3173574.3174156. URL http://dl.acm.org/citation.cfm? doid=3173574.3174156
2. Almada, M.: Human intervention in automated decision-making: Toward the construction of contestable systems. In: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law - ICAIL '19, p. 2–11. ACM Press (2019). DOI 10.1145/3322640.3326699. URL http://dl.acm.org/citation.cfm?doid=3322640.3326699
3. Alvarez, M.: Reasons for Action: Justification, Motivation, Explanation. In: E.N. Zalta (ed.) The Stanford Encyclopedia of Philosophy, winter 2017 edn. Metaphysics Research Lab, Stanford University (2017)
4. Atkinson, K., Baroni, P., Giacomin, M., Hunter, A., Prakken, H., Reed, C., Simari, G., Thimm, M., Villata, S.: Towards artificial argumentation. AI Magazine **38**(3), 25–36 (2017). DOI 10.1609/aimag.v38i3.2704. URL https://www.aaai.org/ojs/index.php/aimagazine/article/view/2704
5. Bex, F., Walton, D.: Combining explanation and argumentation in dialogue. Argument and Computation **7**(1), 55–68 (2011)
6. Binns, R.: Algorithmic accountability and public reason. Philos. Technol. **31**, 543–556 (2018). URL https://doi.org/10.1007/s13347-017-0263-5
7. Biran, O., Cotton, C.: Explanation and justification in machine learning: A survey. In: IJCAI-17 workshop on explainable AI (XAI), vol. 8, pp. 8–13 (2017)
8. Biran, O., McKeown, K.: Justification narratives for individual classifications. In: Proceedings of the AutoML workshop at ICML, vol. 2014, pp. 1–7 (2014)
9. Biran, O., McKeown, K.R.: Human-centric justification of machine learning predictions. In: IJCAI, p. 1461–1467 (2017)
10. Corfield, D.: Varieties of justification in machine learning. Minds and Machines **20**(2), 291–301 (2010). DOI 10.1007/s11023-010-9191-1
11. Diakopoulos, N.: Accountability in algorithmic decision making. Commun. ACM **59**(2), 56–62 (2016). DOI 10.1145/2844110. URL https://doi.org/10.1145/2844110
12. Hickok, M.: Lessons learned from AI ethics principles for future actions. AI and Ethics pp. s43681–020–00008–1 (2020). DOI 10.1007/s43681-020-00008-1
13. Hirsch, T., Merced, K., Narayanan, S., Imel, Z.E., Atkins, D.C.: Designing contestability: Interaction design, machine learning, and mental health. In: Proceedings of the 2017 Conference on Designing Interactive Systems, DIS '17, p. 95–99. Association for Computing Machinery, New York, NY, USA (2017). DOI 10.1145/3064663.3064703. URL https://doi.org/10.1145/3064663.3064703
14. Irving, G., Christiano, P., Amodei, D.: AI safety via debate. arXiv:1805.00899 [cs, stat] (2018). URL http://arxiv.org/abs/1805.00899. ArXiv: 1805.00899
15. Kass, R., Finin, T., et al.: The need for user models in generating expert system explanations. International Journal of Expert Systems **1**(4) (1988)
16. Kim, B.: Interactive and interpretable machine learning models for human machine collaboration. Ph.D. thesis, Massachusetts Institute of Technology (2015)

17. Langley, P.: Explainable, normative, and justified agency. Proceedings of the AAAI Conference on Artificial Intelligence **33**, 9775–9779 (2019). DOI 10.1609/aaai.v33i01.33019775
18. Laugel, T., Lesot, M.J., Marsala, C., Renard, X., Detyniecki, M.: The dangers of post-hoc interpretability: Unjustified counterfactual explanations. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, p. 2801–2807. International Joint Conferences on Artificial Intelligence Organization (2019). DOI 10.24963/ijcai.2019/388. URL https://www.ijcai.org/proceedings/2019/388
19. Lei, T., Barzilay, R., Jaakkola, T.: Rationalizing neural predictions. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, p. 107–117. Association for Computational Linguistics (2016). DOI 10.18653/v1/D16-1011. URL http://aclweb.org/anthology/D16-1011
20. Liao, B., Anderson, M., Anderson, S.L.: Representation, justification, and explanation in a value-driven agent: an argumentation-based approach. AI and Ethics pp. s43681–020–00001–00008 (2020). DOI 10.1007/s43681-020-00001-8
21. Loi, M., Ferrario, A., Viganò, E.: Transparency as design publicity: Explaining and justifying inscrutable algorithms. SSRN Electronic Journal (2019). DOI 10.2139/ssrn.3404040. URL https://www.ssrn.com/ abstract=3404040
22. Madumal, P., Miller, T., Sonenberg, L., Vetere, F.: A grounded interaction protocol for explainable artificial intelligence. In: Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19, p. 1033–1041. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC (2019)
23. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence **267** (2017). DOI 10.1016/j.artint.2018.07.007
24. Miller, T., Howe, P., Sonenberg, L.: Explainable AI: Beware of inmates running the asylum. In: IJCAI-17 Workshop on Explainable AI (XAI), vol. 36 (2017)
25. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D., Gebru, T.: Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19 (2019). DOI 10.1145/3287560.3287596. URL http://dx.doi.org/10.1145/3287560.3287596
26. Mittelstadt, B., Russell, C., Wachter, S.: Explaining explanations in ai. In: Proceedings of the conference on fairness, accountability, and transparency, p. 279–288 (2019)
27. Mueller, S.T., Hoffman, R.R., Clancey, W., Emrey, A., Klein, G.: Explanation in human-ai systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable ai. arXiv preprint arXiv:1902.01876 (2019)
28. Mulligan, D.K., Kluttz, D., Kohli, N.: Shaping our tools: Contestability as a means to promote responsible algorithmic decision making in the professions. Available at SSRN 3311894 (2019). URL https://ssrn.com/abstract=3311894
29. Ploug, T., Holm, S.: The four dimensions of contestable ai diagnostics - a patient-centric approach to explainable ai. Artificial Intelligence in Medicine **107**, 101901 (2020). DOI 10.1016/j.artmed.2020.101901
30. Swartout, W.R.: Explaining and justifying expert consulting programs. In: Computer-assisted medical decision making, pp. 254–271. Springer (1985)
31. Sørmo, F., Cassens, J., Aamodt, A.: Explanation in case-based reasoning–perspectives and goals. Artificial Intelligence Review **24**(2), 109–143 (2005). DOI 10.1007/s10462-005-4607-7
32. Timmons, M.: Moral Theory. Rowman and Littlefield Publishers (2013)